The goal of this problem is to formulate and solve a classification problem, in which we would like to know whether an individual is diabetic (Class 1) or non-diabetic (Class 0). Diabetes is one of the most commonly diseases in the U.S. Among the main factors causing diabetes, two main characteristics include:

1. Elevated body mass index (BMI), noted as variable d_1

2. Family history of diabetes, noted as variable d_2

For this reason, we have obtained data from three individuals, "samples" as they are called in machine learning, for which we have measured their BMI and the number of relatives suffering from diabetes.
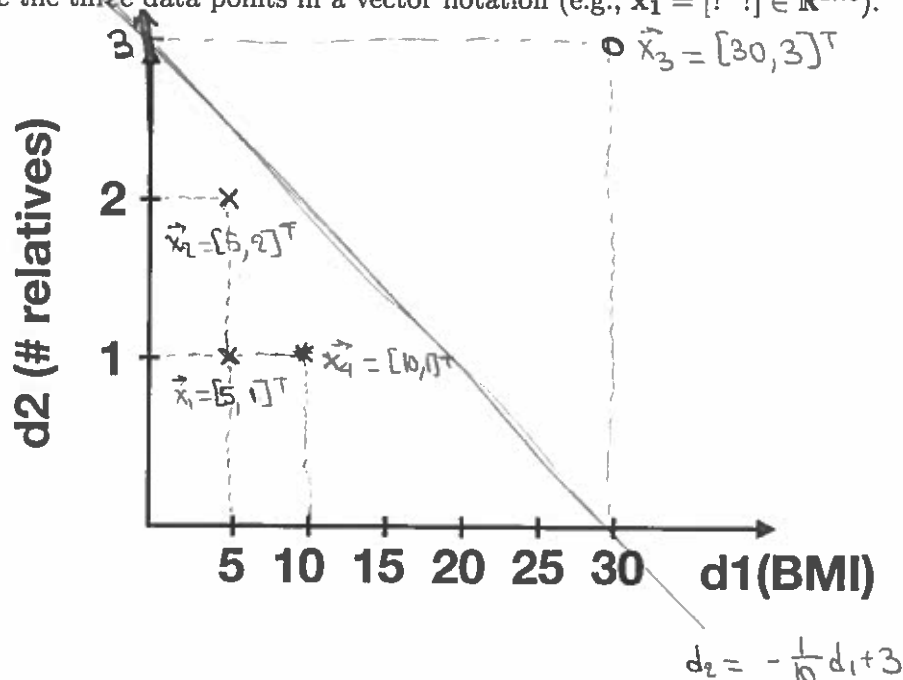
1. Sample 1 (x_1) is non-diabetic, has a BMI of 5 and 1 relative with diabetes

2. Sample 2 (x_2) is non-diabetic, has a BMI of 5 and 2 relatives with diabetes

3. Sample 3 (x_3) is diabetic, has a BMI of 30 and 3 relatives with diabetes

In this problem, we are going to see:

- How to express the aforementioned information in a vector notation (Part 1)

- How to obtain a mathematical expression to classify between the diabetics and non-diabetics (Part 2)

- How to obtain a decision for an "unknown" sample, for which we do not whether he/she is diabetic or not (Parts 3 and 4)

Part 1: Forming a 2D vector space from the data

Consider the above data in a 2D space, where the horizontal axis d_1 is BMI and vertical axis d_2 is the number of relatives with diabetes. Please plot the three data points x_1, x_2, and x_3 and write the three data points in a vector notation (e.g., $x_1 = [?\ ?] \in \mathbb{R}^{2\times 1}$).

$$\vec{x_3} = [30,3]^T$$
$$\vec{x_2} = [5,2]^T$$
$$\vec{x_4} = [10,1]^T$$
$$\vec{x_1} = [5,1]^T$$
$$d_2 = -\tfrac{1}{10}d_1 + 3$$

Part 2: Forming a linear decision boundary

We would like to find a line that separates diabetic from non-diabetic samples. This is called "decision boundary," and since we are interested in a line, we are looking at a "linear decision boundary."

Consider that the line passing from $[30, 0]$ and $[0, 3]$ as a decision boundary for classifying between diabetic and non-diabetic individuals.

(2.a) Please plot the line in the above figure and find its equation.
Hint: You will need to find the equation of the line in terms of d_1 and d_2.

$$d_2 = \underbrace{-\frac{1}{10} d_1}_{\text{slope}} + \underbrace{3}_{\text{y-intercept}}$$

(2.b) Using the above line, form a classification function, i.e., $f(d_1, d_2)$ which will decide whether a new sample $\mathbf{x} = [d_1, d_2]^T$ belongs to Class 0 (non-diabetic) or Class 1 (diabetic).
Hint: You will need to express f with a rule-based equation (e.g., if $f(d_1, d_2) > 0$ then Class 1, else ...)

A sample $\hat{x} = [d_1, d_2]^T$ belongs to Class 1, if it is on the right

of the line $d_2 = -\frac{1}{10} d_1 + 3$, therefore it meets the condition:

$$d_2 > -\frac{1}{10} d_1 + 3 \implies -3 + \frac{1}{10} d_1 + d_2 > 0$$

Similarly, a sample $\hat{x}$ belongs to Class 0, if it lies on the left of

the line, therefore, it should meet the condition:

$$d_2 < -\frac{1}{10} d_1 + 3 \implies -3 + \frac{1}{10} d_1 + d_2 < 0$$

Therefore the classification function should be:

$$\begin{cases} \text{if} \quad f(d_1, d_2) = -3 + \frac{1}{10} d_1 + d_2 > 0, \text{ then } \hat{x} = [d_1, d_2]^T \in \text{Class 1} \\ \text{if} \quad f(d_1, d_2) = -3 + \frac{1}{10} d_1 + d_2 < 0, \text{ then } \hat{x} = [d_1, d_2]^T \in \text{Class 0} \end{cases}$$

(2.c) Express the three given data points in a matrix notation, such that $\mathbf{D} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \in \mathbb{R}^{3 \times 3}$,

by substituting the values of x_1, x_2, and x_3 in the above equation. Multiply $\mathbf{D}$ with the weight $\mathbf{w} = [w_0, w_1, w_3]^T = [-3, \frac{1}{10}, 1]^T \in \mathbb{R}^{3 \times 1}$. What is the final dimension of $\mathbf{D} \cdot \mathbf{w}$? What do you observe based on question **(2.b)**?

$$\mathbf{D} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 30 & 3 \end{bmatrix}$$

$$D \cdot w = \begin{bmatrix} 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 30 & 3 \end{bmatrix} \begin{bmatrix} -3 \\ 1/10 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 + \frac{1}{2} + 1 \\ -3 + \frac{1}{2} + 2 \\ -3 + 3 + 3 \end{bmatrix} = \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{2} \\ 3 \end{bmatrix} \begin{matrix} <0 \\ <0 \\ >0 \end{matrix}$$

We observe that $D\vec{w}$ is the vector of the decision rule from notation

question (2.b), since for a given $\vec{x} = [d_1, d_2]^T$, we have

$$\underbrace{[1 \; d_1 \; d_2]}_{[1 \; \vec{x}]} \underbrace{\begin{bmatrix} -3 \\ 1/10 \\ 1 \end{bmatrix}}_{\vec{w}} = -3 + \frac{1}{10}d_1 + d_2 = f(d_1, d_2)$$

(2.d) Now assume that you would like to find whether a person x_4 whose BMI is 10 and has 1 relative with diabetes, is a diabetic or not. How would you write x_4 in a vector notation? How would you classify this person based on the rule in question (2.b)?

$$\vec{x_4} = [10, 1]^T$$

$$f(\vec{x_4}) = -3 + \frac{1}{10} 10 + 1 = -1 < 0$$

therefore $\vec{x_4}$ belongs to Class 0 (non-diabetic)

Part 3: 1-Nearest Neighbor with Euclidean distance

(3.a) Compute the Euclidean distance ($l2-$norm) between sample x_4 and samples x_1, x_2, x_3.
Hint: If $a = [a_1, a_2]^T$ and $b = [b_1, b_2]^T$, then $l_2(a, b) = \|a - b\|_2 = \sqrt{(|a_1 - b_1|^2 + |a_2 - b_2|^2)}$.

$$x_1 = [5,1]^T \;,\; \vec{x}_2 = [5,2]^T \;,\; \vec{x}_3 = [30,3]^T \;,\; \vec{x}_4 = [10,1]^T$$

$$d(\vec{x}_1, \vec{x}_4) = \sqrt{(5-10)^2 + (1-1)^2} = 5$$

$$d(\vec{x}_2, \vec{x}_4) = \sqrt{(5-10)^2 + (2-1)^2} = \sqrt{26}$$

$$d(\vec{x}_3, \vec{x}_4) = \sqrt{(30-10)^2 + (3-1)^2} = \sqrt{20^2 + 2^2} = \sqrt{404}$$

(3.b) Which sample is closest to x_4 based on the Euclidean distance?

$$d(\vec{x}_1, \vec{x}_4) < d(\vec{x}_2, \vec{x}_4) < d(\vec{x}_3, \vec{x}_4)$$

therefore $\vec{x}_4$ is closest to $\vec{x}_1$ based on the Euclidean distance.

(3.c) If we were to classify x_4 based on the class of its closest sample according to the Euclidean distance, would it come from a diabetic or a non-diabetic person?

Since $\vec{x}_1$ is non-diabetic, and $\vec{x}_4$ is closest to $\vec{x}_1$ based on the Euclidean distance, then $\vec{x}_4$ is non-diabetic.

4

Part 4: 1-Nearest Neighbor with cosine similarity

(4.a) Compute the cosine similarity between sample x_4 and samples x_1, x_2, x_3. Hint: If $a = [a_1, a_2]^T$ and $b = [b_1, b_2]^T$, then $cos(a, b) = \frac{(a,b)}{\|a\|_2 \|b\|_2} = \frac{a_1 \cdot b_1 + a_2 \cdot b_2}{\sqrt{(|a_1|^2 + |a_2|^2)} \sqrt{(|b_1|^2 + |b_2|^2)}}$.

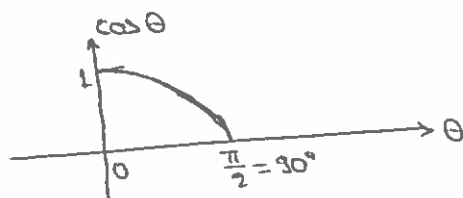$$\vec{x_1} = [5,1]^T, \quad \vec{x_2} = [5,2]^T, \quad \vec{x_3} = [30,3]^T, \quad \vec{x_4} = [10,1]^T$$

$$cos(\vec{x_1}, \vec{x_4}) = \frac{5 \times 10 + 1 \times 1}{\sqrt{(5^2 + 1^2)(10^2 + 1^2)}} \cong 0.99 \implies \theta(\vec{x_1}, \vec{x_4}) \simeq 0.14\pi \simeq 44°$$

$$cos(\vec{x_2}, \vec{x_4}) = \frac{5 \times 10 + 2 \times 1}{\sqrt{(5^2 + 2^2)(10^2 + 1^2)}} \cong 0.96 \implies \theta(\vec{x_2}, \vec{x_4}) \simeq 0.28\pi \simeq 88°$$

$$cos(\vec{x_3}, \vec{x_4}) = \frac{30 \times 10 + 3 \times 1}{\sqrt{(30^2 + 3^2)(10^2 + 1^2)}} = \frac{303}{\sqrt{909 \times 101}} = \frac{303}{\sqrt{101^2 \times}} = 1$$

$$\implies \theta(\vec{x_3}, \vec{x_4}) = 0 \simeq 0°$$

In fact, we see that $\vec{x_3}$ and $\vec{x_4}$ lie in the same line starting from o.

(4.b) Which sample is closest to x_4 based on the cosine distance?

Based on the above, $\vec{x_4}$ forms a 0° angle with $\vec{x_3}$, while it forms a 44° and 88° angle with $\vec{x_1}$ and $\vec{x_2}$. Therefore based on the cosine distance $\vec{x_4}$ is closest to $\vec{x_3}$.

(4.c) If we were to classify x_4 based on the class of its closest sample according to the cosine distance, would it come from a diabetic or a non-diabetic person? What do you observe compared to question (3.c)?

Sample $\vec{x_3}$ belongs to the diabetic class, therefore based on the cosine similarity, sample $\vec{x_4}$ is diabetic. This contradicts the decision based on the Euclidean distance (question 3.c), which indicates that the distance metric matters when comparing samples in the $\mathbb{R}^5$ space.